Information-related complexity: a problem-oriented approach

E. Perevalov[*] and D. Grace[†]

Lehigh University

Bethlehem, PA

A general notion of information-related complexity applicable to both natural and man-made systems is proposed. The overall approach is to explicitly consider a rational agent performing a certain task with a quantifiable degree of success. The complexity is defined as the minimum (quasi-)quantity of information that's necessary to complete the task to the given extent – measured by the corresponding loss. The complexity so defined is shown to generalize the existing notion of statistical complexity when the system in question can be described by a discrete-time stochastic process. The proposed definition also applies, in particular, to optimization and decision making problems under uncertainty in which case it gives the agent a useful measure of the problem's "susceptibility" to additional information and allows for an estimation of the potential value of the latter.

---

[*] E-mail: eup2@lehigh.edu

[†] E-mail: dpg3@lehigh.edu

I. INTRODUCTION

The general field of complex systems science is relatively new and has been developing rather rapidly in the last three decades. One of the central notions it revolves around is – understandably – that of complexity. The desire to define complexity before attempting to study it is indeed rather natural and logical in this context. When the field of complex system science began taking its shape several notions of complexity were already in use. The most notable of them was the Kolmogorov descriptional complexity [1, 2] defined as a shortest length of a computer program needed to reproduce the given object (in the simplest case represented by a string of symbols from a given alphabet). It was quickly noted, however, that the Kolmogorov complexity is rather a measure of *randomness* as opposed to complexity understood as the amount of appropriately defined "internal structure" possessed by the object [3]. It was also realized that a proper measure of complexity should be probabilistic since complexity as it is encountered in natural sciences and engineering is inherently linked to some sort of *uncertainty*. Following these realizations, several useful measures of complexity were developed, including set complexity, true measure complexity, effective measure complexity [4], logical depth [5, 6], statistical complexity [7], sophistication [8]. Among these measures, statistical complexity was the measure foundations and applications of which were later developed in most detail (see [9]). For this reason, we will use it later as a "benchmark" complexity measure for systems that can be described by discrete-time stochastic processes.

The main goal of the present work is to put the problem of complexity measures in an "engineering" problem-oriented context and, in particular, to demonstrate that a generalization of statistical complexity measure arises naturally when the proposed "engineering" complexity notion is applied to the case of a prediction task for discrete-time stochastic processes. The central figure of the proposed approach is an *agent* whose goal is to perform a certain *task*. Typical examples of tasks are control, decision making, optimization, prediction. We will also refer to the such tasks as *problems*. From a general practical point of view, a task can be made complex by several of its aspects. It can be, for example, computationally complex (like a large scale deterministic optimization problem). It can be operationally complex (like that of building a complicated object given full instructions). We will not deal with such complexity types. Instead, we concentrate on *information*-related complexity, the measure of which is an appropriately defined quantity of some relevant information needed for completing the task with a given (also appropriately measured) degree of success. This implies that the task is initially associated with some degree of uncertainty

that can be fully or partially removed by acquiring the appropriate information.

The rest of the paper is organized as follows. In Section II, the general definition of task complexity is given. Section III applies considers the optimization under uncertainty (stochastic optimization) as a task and derives the quasi-information complexity for optimization. Section IV considers the task of prediction of stochastic processes output and makes a connection to statistical complexity introduced in [7] which turns out to be just the task complexity taken at zero value of the argument. Finally, Section V summarizes the contributions of the present work and presents a brief informal discussion.

II. GENERAL FORMULATION: INFORMATION-RELATED TASK COMPLEXITY

An agent is trying to accomplish a certain task and can, in principle, be helped by appropriate additional information. The quality (or degree) of the task accomplishment is assumed to be quantifiable by a real number and thus can be compared between any two instances of the task accomplishment. Respectively, a notion of real-valued *loss* can be defined as (the absolute value of) the difference between the degree of the task accomplishment in the given instance and the best possible one. The latter has to be defined appropriately given the specific task at hand as examples given later in the paper illustrate.

Besides loss, any information $\mathfrak{I}$ that the agent can acquire is also assumed to possess some real-valued measure $I(\mathfrak{I})$ of its effective quantity which we will refer to as *quasi-quantity* to distinguish it from measures of "pure" quantity such as Shannon entropy. The appropriate form of quasi-quantity measure $I(\mathfrak{I})$ has to be determined by considering the specific nature of the task in hand to appropriately reflect its essential features. In particular, the quasi-quantity measure may coincide with a traditional "pure" quantity measure for some tasks. The agent is also assumed to be capable of making optimal use of the acquired information for the purpose of minimizing the loss. Given the latter assumption, we can introduce the notion of the loss $L(\mathfrak{I})$ of the specific information $\mathfrak{I}$ available to the agent.

It is natural to assume that, if the measure of information quantity adapted in the given context is adequate, then smaller loss values can in general be obtained for larger values of the information quasi-quantity. Thus, to reduce the loss more, larger quantities of information would in general be needed. On the other hand, it is clear that, in general, different "pieces" of information of equal quasi-quantity would lead to different values of loss reduction. Therefore, if loss reduction at a minimum information "cost" is the agent's goal, it makes sense to search for the specific information

that would allow the agent to achieve that goal. It is also rather natural to use the initial value of the loss $L_0 \equiv L(\emptyset)$ that the agent can obtain in the absence of any (additional) information as a reference point for the reduced loss. Specifically, one can use the ratio $\gamma = \frac{L(\mathcal{I})}{L_0}$ as a natural dimensionless measure of the achieved loss reduction. The lowest information quasi-quantity needed to achieve such a loss reduction can be associated with the proper *quasi-information complexity* C_γ of the task in question. This leads us to the following definition.

$$C_\gamma = \min_{\{\mathcal{I}: L(\mathcal{I}) \leq \gamma L_0\}} I(\mathcal{I}). \tag{1}$$

Note that the minimization in (1) is performed over all possible "kinds", or "pieces" of information that – if used optimally by the agent – would yield the value of loss not exceeding the given fraction γ of the original value L_0. Note also that the specific interpretations of the loss $L(\cdot)$ and quantity $I(\cdot)$ are intentionally left open by the definition (1). The understanding is that they have to be made more concrete by taking the "nature" and the details of the system in question into account. In the next two sections, we consider two more specific examples that illustrate the use of the proposed definition.

III.   OPTIMIZATION

Consider a typical "engineering" case of optimization under uncertainty, with the expected value of the objective function being the main optimization criterion (stochastic optimization formulation). The overall task can be concisely written as follows.

$$\min_{x \in X} \int_\Omega f(\omega, x) P(d\omega), \tag{2}$$

where $X \subset \mathcal{D}$ is the set of all feasible solutions, i.e. the set satisfying all (deterministic) constraints that may be present in the problem formulation (with $\mathcal{D}$ being the space to which all solutions belong like a suitable Euclidean space); Ω is the base space of possible values of input data parameters that are not known with certainty often referred to as a *parameter space*; P is a probability measure on $(\Omega, \mathcal{F})$ (where $\mathcal{F}$ is a suitable sigma-algebra) that describes the initial state of information available to the agent; $f: \Omega \times \mathcal{D} \to \overline{\mathbb{R}}$ is a function integrable on Ω for each $x \in X$ describing the specific optimization objective.

The value of the initial loss L_0 then becomes simply

$$L_0 = L(P) = \int_\Omega (f(\omega, x_P^*) - f(\omega, x_\omega^*)) P(d\omega) \tag{3}$$

here x_P^* is a solution of (2) and x_ω^* is a solution of $\min_{x \in X} f(\omega, x)$ for the given ω of the parameter space Ω.

As it has been argued in [10] and [11], in the general context of problem (2), additional information can be obtained by the agent by means of querying *information sources*. This can be accomplished by the agent via formulating and asking *questions* and receiving corresponding *answers* from the source(s). It has been further argued that the appropriate meaning of information quasi-quantity is supplied by *pseudoenergy* that provides real-valued measure for both *question difficulty* and *answer depth* functionals. The reason pseudoenergy is appropriate in the given context is that it takes into account the source's *knowledge structure* and therefore allows the agent to correctly select optimal "pieces" of information to maximize their effect on the given problem.

The role of "pieces" of information themselves is played by questions which can be associated, following the earlier developments of [12–14] and [15, 16], with *partitions* of the problem parameter space Ω. The corresponding quasi-quantity is then given by the question difficulty functional $G(\Omega, \mathbf{C}, P)$ where the partition $\mathbf{C} = \{C_1, \ldots, C_r\}$ is a collection of subsets of Ω such that $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^r C_i = \Omega$. The loss $L(\mathbf{C})$ corresponding to question $\mathbf{C}$ can be interpreted as the minimum loss that an optimizing agent can obtain upon reception of a perfectly accurate answer to question $\mathbf{C}$ can be easily seen to take the form

$$L(\mathbf{C}, P) = \sum_{j=1}^r P(C_j) \int_{C_j} (f(\omega, x_{P_j}^*) - f(\omega, x_\omega^*)) dP_{C_j}(\omega), \tag{4}$$

where P_{C_j} is the conditional measure corresponding to the subset $C_j \in \mathbf{C}$. It is straightforward to show that $L(\mathbf{C}) \leq L_0$. Correspondingly, the informational complexity C_γ of the optimization task (2) takes the form

$$C_\gamma = \min_{\{\mathbf{C}: L(\mathbf{C}, P) \leq \gamma L(P)\}} G(\Omega, \mathbf{C}, P), \tag{5}$$

where minimization in (5) is over all possible partitions of Ω. The form of informational complexity (5) assumes a specific source knowledge structure which is encoded in the form of question difficulty functional $G(\cdot)$. In particular, as was shown in [10], in an isotropic (but in general non-uniform) case, the functional $G(\cdot)$ depends besides the "initial" probability measure P, on an integrable function $u \colon \Omega \to \mathbb{R}_+$ on the parameter space Ω:

$$G(\Omega, \mathbf{C}, P) = \frac{\sum_{j=1}^r u(C_j) P(C_j) \log \frac{1}{P(C_j)}}{\sum_{j=1}^r P(C_j)},$$

where $u(C_j) = \frac{\int_{C_j} u(\omega)\, dP(\omega)}{P(C_j)}$. If the knowledge structure is unknown, it appears reasonable to assume the most symmetric case (uniform and isotropic) in order to obtain a "benchmark" (or

TABLE I. Quasi-information complexity C_γ for some values of γ for Example 1.

γ	0.2	0.1	0.05	0.01	0
C_γ	1.2	1.7	2.2	3.8	∞

"universal") task complexity. In this case, the question difficulty reduces to Shannon entropy $H(P(\mathbf{C})) \equiv H(P(C_1), \ldots, P(C_r))$ of the discrete probability distribution induced by the partition $\mathbf{C}$, and the informational task complexity C_γ becomes simply

$$C_\gamma = \min_{\{\mathbf{C}:L(\mathbf{C},P)\leq\gamma L(P)\}} H(P(\mathbf{C})). \tag{6}$$

A. Example

Consider the stochastic optimization problem of the form

$$\text{maximize} \quad x_1 + \omega x_2$$
$$\text{subject to} \quad x_1^2 + x_2^2 \leq 1$$
$$x_1 \geq 0, \ x_2 \geq 0$$

where the only uncertain parameter is ω which has a uniform distribution on the interval $\Omega = [0,1]$. The situation is depicted in Fig. 1 and is referred to as Example 1 in what follows. It is straightforward to show that for a partition $\mathbf{C}$ of Ω into r subsets of equal measure, the loss (4) takes the form

$$L(\mathbf{C},P) = \int_0^1 \sqrt{1+\omega^2}\, d\omega - \frac{1}{r}\sum_{i=1}^r \sqrt{1 + \left(\frac{2i-1}{2r}\right)^2}.$$

On the other hand, if one uses (in the absence of information about possible information sources and their knowledge structure) the uniform isotropic model, the difficulty functional reduces to Shannon entropy and $G(\Omega, \mathbf{C}, P) = H(P(\mathbf{C})) = \log r$. The relationship between the information quasi-quantity $H(P(\mathbf{C})$ and the corresponding loss (rescaled so that $L_0 = 0$) is shown in Fig. 2. The task complexity C_γ for some values of the parameter γ are shown in Table I. In particular, the "full" task complexity C_0 for this example is infinite.

Suppose now that the parameter space Ω and the initial measure P are the same but the task is somewhat different. Specifically, the stochastic optimization problem has the linear form (see
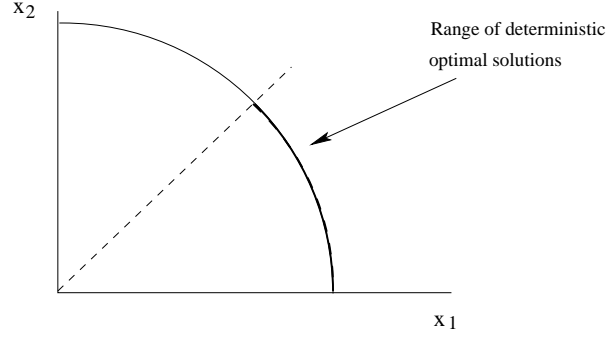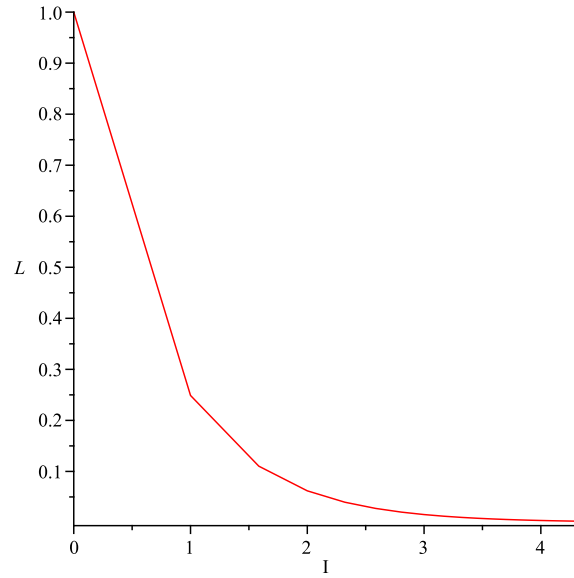
FIG. 1. Example 1.

FIG. 2. Loss vs. information quasi-quantity for example 1.

Fig. 3 for an illustration). We refer to this example as Example 2.

$$\text{maximize} \quad x_1 + \omega x_2$$

$$\text{subject to} \quad x_1 + (\sqrt{2} - 1)x_2 \leq 1$$

$$(\sqrt{2} - 1)x_1 + x_2 \leq 1$$

$$x_1 \geq 0, \; x_2 \geq 0$$

In this case, as it is easy to see, it is sufficient to consider questions corresponding to partitions of Ω into two subsets only, in order to fully get rid of the agent uncertainty with respect to the optimal solution – assuming that perfectly accurate answers to such questions are available. Specifically, consider partitions of Ω of the form $\mathbf{C} = \{[0, a], (a, 1]\}$ where $0 \leq a \leq \omega_c = \sqrt{2} - 1$. Then the loss

TABLE II. Quasi-information complexity C_γ for some values of γ for Example 2.

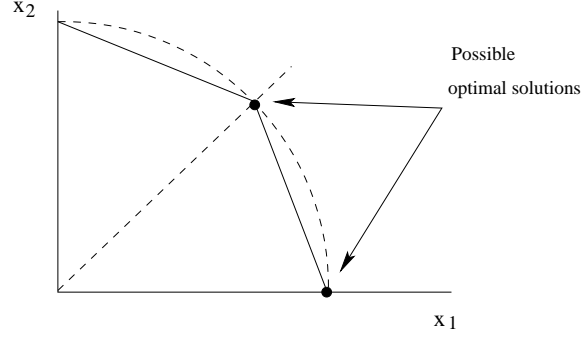| γ	0.2	0.1	0.05	0.01	0
C_γ	0.77	0.85	0.90	0.95	0.98

FIG. 3. Example 2.

(4) takes the form

$$L(\mathbf{C}, P) = \frac{\sqrt{2} - 1}{\sqrt{2}}(\omega_c - a) - \frac{1}{2\sqrt{2}}(\omega_c^2 - a^2),$$

and the information quasi-quantity becomes

$$G(\Omega, \mathbf{C}, P) = H(P(\mathbf{C})) = H(a) = -a \log a - (1 - a) \log(1 - a).$$

The relationship between them is shown in Fig. 4. The task complexities for some values of γ are shown in Table II. One can see that the "full" complexity C_0 for this example is finite.

Let us now make some observations (and careful generalizations).

- Even if the "uncertainty structure" (the parameter space Ω and the measure P) is precisely the same, the task complexities may be different (and even very different) if the agent is solving different problems. In the two examples just discussed, for instance, the complexity C_0 is infinite for Example 1 while it is and finite (and rather small) for Example 2.

- The "full" complexity C_0 typically does not tell the whole story about the "information content" of a problem.

- The whole curve C_γ is in principle needed to obtain a picture of the problem's "information susceptibility", or, equivalently, the ability of additional information to improve the solution quality.
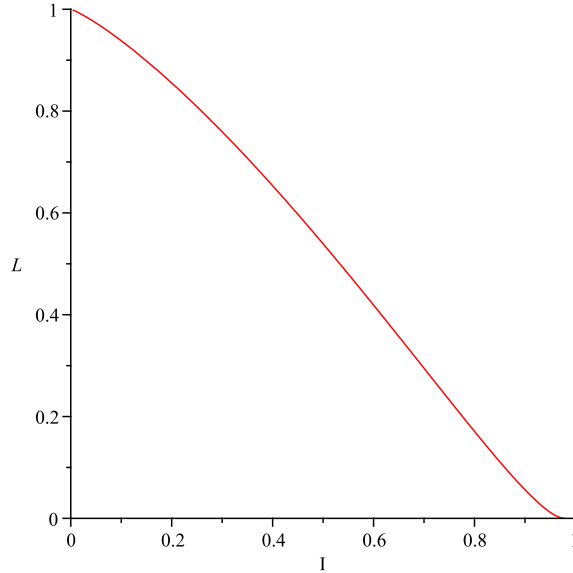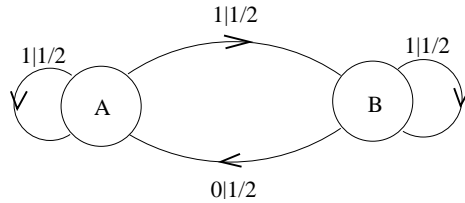
FIG. 4. Loss vs. information quasi-quantity for example 2.

FIG. 5. Generation of observed symbols.

- The knowledge of the curve C_γ, coupled with that of available additional information sources and their knowledge structure (see [10, 11] for a discussion of the latter), allows the agent to evaluate the effectiveness of additional information for the particular problem and its corresponding value.

IV. STOCHASTIC PROCESSES AND STATISTICAL COMPLEXITY

Consider now an application from a different domain: the case of discrete-time stochastic processes. To make the discussion more concrete, let us consider a specific simple example (referred to as Example A in what follows) which is borrowed from [17]. Suppose the unobserved symbols A and B are generated randomly with equal probability. Symbols 0 and 1 – that are observed – correspond to pairs of A and B as follows: $\{AA, AB, BB\} \to 1$, $\{BA\} \to 0$ (see Fig. 5 for an illustration). For example, if $AABAB$ was generated, the observer will see 1101.

The concept of ϵ-machines was introduced in [7] and fully developed in [9] to describe the
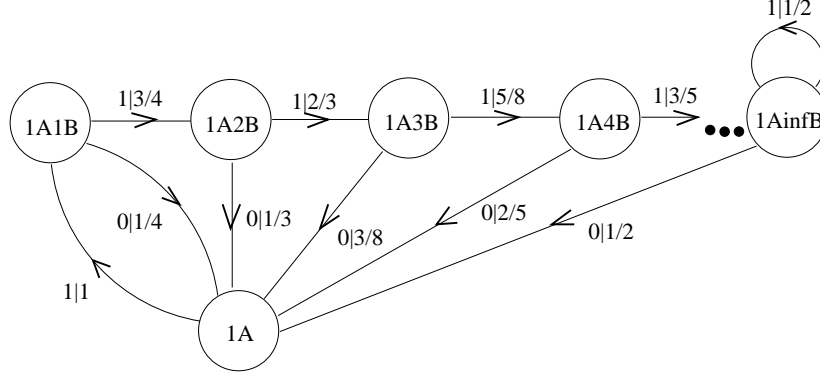
FIG. 6. Exact ϵ-machine for Example A.

intuitive (and practical) notion of complexity as the amount of underlying "structure" in the system. ϵ-machines consist of casual states and transitions between them (see [9] for a detailed exposition). Each time a transition happens, an observed symbols is output. They can be thought of as functions on Markov chains. Causal states are equivalence classes of previous histories: given two histories belonging to the same causal state S the conditional distribution of future string of bits is the same. Thus ϵ-machines are *deterministic* meaning that given a causal state and an output symbol, the next causal state is uniquely determined. ϵ-machines can in principle be reconstructed from binary tree representation of the system output. Considering longer symbol sequences (words) allows for distinguishing more distinct causal states (which in turn requires longer output sample sizes). It can be shown that, for our example, the full ϵ-machine has the form shown in Fig. 6. The transition probability for this ϵ-machine have the form $\Pr(1AiB \to 1A(i+1)B) = \frac{i+2}{2(i+1)}$, $\Pr(1AiB \to 1A) = \frac{i}{2(i+1)}$, and $\Pr(1A \to 1A1B) = 1$. Any transition to the state $1A$ outputs a symbol 0; all other transitions output 1.

Given an ϵ-machine $\mathcal{M}$ for the system, the *statistical complexity* C_μ is defined as the minimum expected amount of information that needs to be kept track of for optimal prediction of the next bit (which in general is imperfect):

$$C_\mu(\mathcal{M}) = -\sum_{S \in \mathcal{M}} P(S) \log P(S), \tag{7}$$

where $P(S)$ is the (steady state) probability of the causal state S. The *topological complexity* is a related quantity that reflects just the number of causal states in the ϵ-machine:

$$C_\tau = \log |\mathcal{M}|.$$

For our example, it is easy to see that $P(1AiB) = \frac{i+1}{2^{i+2}}$, $i = 1, 2 \ldots$ and $P(1A) = 1/4$. Thus $C_\mu(\mathcal{M}) = 2.71$. Note that the topological complexity $C_\tau = \log |\mathcal{M}|$ of this system is infinite.

The *entropy rate* for an ϵ-machine (and the corresponding dynamic system) can be found as

$$h_\mu(\mathcal{M}) = -\sum_{S\in\mathcal{M}} P(S) \sum_{s\in\mathcal{A}} P_\mathcal{M}(s|S) \log P_\mathcal{M}(s|S), \tag{8}$$

where $\mathcal{A}$ is the alphabet of output symbols ($\mathcal{A} = \{0,1\}$ in this example). The entropy rate can be thought of as the measure of uncertainty per symbol that cannot be removed by finding the inherent structure of the system. For our example, the entropy rate is found to be equal to 0.678. We will also need the notion of *relative entropy rate* $h_\mu(\mathcal{M}|\mathcal{M}')$ of ϵ-machine $\mathcal{M}$ with respect to another ϵ-machine [18] $\mathcal{M}'$. Let $g\colon \mathcal{M}' \to \mathcal{M}$ be a map between states of these two ϵ-machines. Then the relative entropy rate $h_\mu(\mathcal{M}|\mathcal{M}')$ is defined as

$$h_\mu(\mathcal{M}|\mathcal{M}') = -\sum_{S'\in\mathcal{M}'} P_{\mathcal{M}'}(S') \sum_{s\in\mathcal{A}} P_{\mathcal{M}'}(s|S') \log P_\mathcal{M}(s|g(S')).$$

It is straightforward to show that

$$h_\mu(\mathcal{M}|\mathcal{M}') = h_\mu(\mathcal{M}') + \sum_{S'\in\mathcal{M}'} P_{\mathcal{M}'}(S') KL(P_{\mathcal{M}'}(S')||P_\mathcal{M}(S')), \tag{9}$$

where $KL(P_{\mathcal{M}'}(S')||P_\mathcal{M}(S')) = \sum_{s\in\mathcal{A}} P_{\mathcal{M}'}(s|S') \log \frac{P_{\mathcal{M}'}(s|S')}{P_\mathcal{M}(s|g(S'))}$ is the corresponding Kullback-Liebler distance.

To make a link between task complexity and statistical complexity, we need to decide on the suitable task. It appears reasonable, since in this case we want to study the complexity of our system "in its own right", the logical choice for the task is that of suitably defined *prediction*. We also need to define the corresponding loss in a logical way. Several choices are in principle possible here, with *quadratic* and *logarithmic* loss being the two popular ones in prediction algorithms literature. We will use the latter of the two. Specifically, we will use the following single bit loss function

$$l(s,\sigma) = \begin{cases} -\log\sigma, & \text{if } s = 1 \\ -\log(1-\sigma), & \text{if } s = 0 \end{cases},$$

where $0 \le \sigma \le 1$ is the prediction. It's easy to see that if $p = \Pr(s = 1)$ then

$$\min_\sigma E[l(s,\sigma)] = H(p) \tag{10}$$

with the minimum being achieved for $\sigma = p$. Furthermore, one can show that

$$E[l(s,\sigma)] = H(p) + KL(p||\sigma), \tag{11}$$

where $KL(p||\sigma) = p \log \frac{p}{\sigma} + (1 - p) \log \frac{1-p}{1-\sigma}$ is the Kullback-Liebler distance between distributions $(p, 1 - p)$ and $(\sigma, 1 - \sigma)$.

Comparing (10) and (8) we can see that the entropy rate for the given ϵ-machine represents the minimum possible long-term (i.e. expected) logarithmic prediction loss. In other words, this is the minimum prediction loss that can be obtained by an optimally acting agent in possession of the complete information about the system. It is reasonable to identify the information $\mathcal{I}$ from the general definition (1) with some ϵ-machine [19] describing the system's output. Depending on the particular ϵ-machine available to him, the agent may possess complete or partial information. Let us denote the "full" ϵ-machine corresponding to complete information by $\mathcal{M}^*$. Since the logarithmic loss cannot be reduced below that obtained from $\mathcal{M}^*$, it makes sense to set

$$L(\mathcal{M}^*) = 0 \tag{12}$$

and interpret $L(\mathcal{M})$ for any other ϵ-machine as the expected logarithmic loss in excess of $h_\mu(\mathcal{M}^*)$. One can easily show that the expected logarithmic prediction loss for a system functioning according to the "full" ϵ-machine $\mathcal{M}^*$ but predicted optimally with the help of another ϵ-machine $\mathcal{M}$ is equal to the relative entropy rate $h_\mu(\mathcal{M}|\mathcal{M}^*)$. Using the adapted loss definition (12), we have to subtract the minimum (non-removable) logarithmic expected loss $h_\mu(\mathcal{M}^*)$ to arrive at the loss corresponding to $\mathcal{M}$. Taking into account the relationship (9), we arrive at

$$L(\mathcal{M}) = h_\mu(\mathcal{M}|\mathcal{M}^*) - h_\mu(\mathcal{M}^*) = \sum_{S \in \mathcal{M}^*} P_{\mathcal{M}^*}(S) KL(P_{\mathcal{M}^*}(S)||P_\mathcal{M}(S)). \tag{13}$$

The information quasi-quantity corresponding to the ϵ-machine $\mathcal{M}$ can be defined, in the absence of any information about possible sources and their knowledge structure, with the help of the most symmetric model for the latter – uniform and isotropic. This leads, as was mentioned earlier, to the Shannon entropy for the steady state distribution of the corresponding ϵ-machine which is nothing else but the statistical complexity of $\mathcal{M}$:

$$I(\mathcal{M}) = - \sum_{S \in \mathcal{M}} P(S) \log P(S) = C_\mu(\mathcal{M}). \tag{14}$$

In particular, if the information available to the agent is complete, i.e. $\mathcal{M} = \mathcal{M}^*$, then the information quasi-quantity (14) coincides with the system statistical complexity (7) (and the loss the agent can achieve vanishes).

Let us now find (or estimate) the quasi-information complexity for the prediction task for the original example. The general definition (1) applied for this case reads

$$C_\gamma = \min_{\{\mathcal{M}:L(\mathcal{M}) \le \gamma L_0\}} I(\mathcal{M}), \tag{15}$$

since the role of information $\mathcal{I}$ is played by various ϵ-machines $\mathcal{M}$. The original loss L_0 here should be identified with the expected logarithmic prediction loss that can be obtained without any knowledge of the underlying statistical mechanism, i.e. using a partial ϵ-machine with a single state. The minimum in (15) has to be taken over all possible ϵ-machines $\mathcal{M}$ (and corresponding maps $g\colon \mathcal{M}^* \to \mathcal{M}$). Since the full ϵ-machine $\mathcal{M}^*$ consists of *causal* states, it is sufficient to consider partial ϵ-machines with states corresponding to subsets of causal states of $\mathcal{M}^*$. Specifically, let $\kappa = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ be a *transition-consistent* partition of the set $\mathcal{S}$ of states of the ϵ-machine $\mathcal{M}^*$, i.e. such a partition that symbols output by any transition between two states of $\mathcal{M}^*$ are same for all states in the same subset in the partition. An ϵ-machine $\mathcal{M}$ is obtained from any transition-consistent partition κ of the set of states of $\mathcal{M}^*$ by associating a state S_j of $\mathcal{M}$ with the subset $\mathcal{S}_j$ of κ for $j = 1, \ldots, m$. The transition probability $P(S_j \to S_k)$ for $\mathcal{M}$ is then calculated as $\mathrm{Pr}(S_j \to S_k) = \sum_{S' \in \mathcal{S}_j} P(S')\,\mathrm{Pr}(S' \to S'')$, where S'' is the state [20] in subset $\mathcal{S}_k$ with a nonzero transition probability from S'. We denote the resulting ϵ-machine by $\mathcal{M}(\kappa)$. Given this construction, the relation (15) can be rewritten as

$$C_\gamma = \min_{\{\kappa : L(\mathcal{M}(\kappa)) \leq \gamma L_0\}} I(\mathcal{M}(\kappa)), \tag{16}$$

where minimization is over all transition-consistent partitions of the state set of the full ϵ-machine $\mathcal{M}^*$.

Finding the exact minimum in (16) appears to be a rather difficult problem since the number of transition-consistent partitions of the state set of $\mathcal{M}^*$ is in general exponential in the number of casual states in $\mathcal{M}^*$. Therefore here we settle for some reasonable choices of "partial" information $\mathcal{M}$ instead. This way, clearly, we will generally obtain upper bounds on the exact values of C_γ. Fig. 7 shows several such "partial" ϵ-machines that obtain naturally in the process of ϵ-machine reconstruction from sample outputs of the system. Let us denote the "partial" ϵ-machines with $k+1$ nodes by $\mathcal{M}_k$. The map $g\colon \mathcal{M}^* \to \mathcal{M}_k$ for this case has the form $g(1AjB) = 1AjB$ for $j < k$ and $g(1AjB) = 1AkB$ for $j \geq k$. The values of relative entropy and statistical complexity of these ϵ-machines are shown in Table III. The dependence of the loss on the information quasi-quantity can be found from the data of Table III by a simple rescaling (so that $L(\mathcal{M}_1) = 1$). The results are shown in Fig. 8. The quasi-information complexity C_γ for different values of the parameter γ can be read directly from the plot. Some of the resulting values are shown in Table IV.

To consider a different example (called Example B in what follows), let us keep the topology of the exact ϵ-machine for the example just discussed but change the transition probabilities so that the steady state probabilities do not decay as fast "down the line" (see Fig. 9). Specifically,

TABLE III. Relative entropy rates and statistical complexities for sequential approximations to the exact ϵ-machine for Example A.

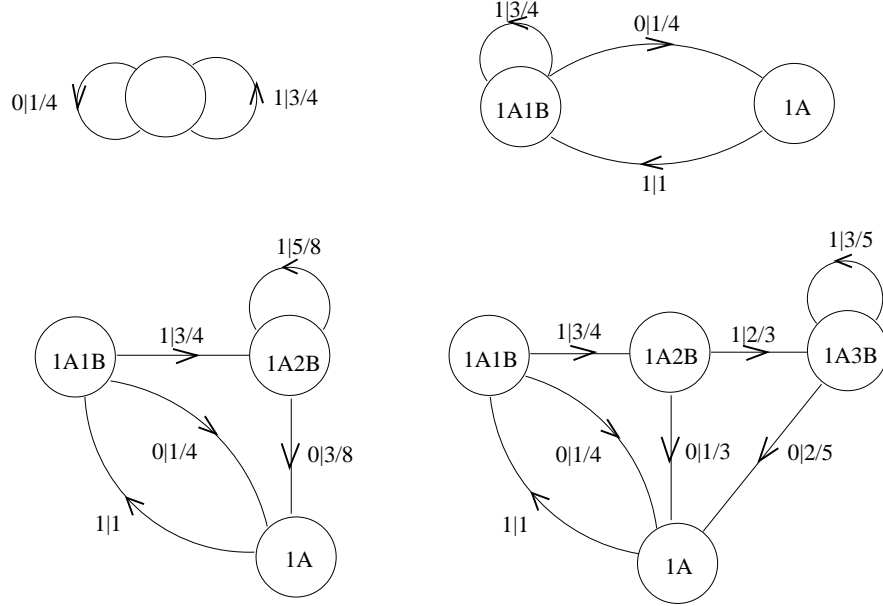| k	0	1	2	3	∞
$h_\mu(\mathcal{M}_k\|\mathcal{M}^*)$	0.811	0.708	0.683	0.679	0.678
$C_\mu(\mathcal{M}_k)$	0	1.47	1.97	2.28	2.71

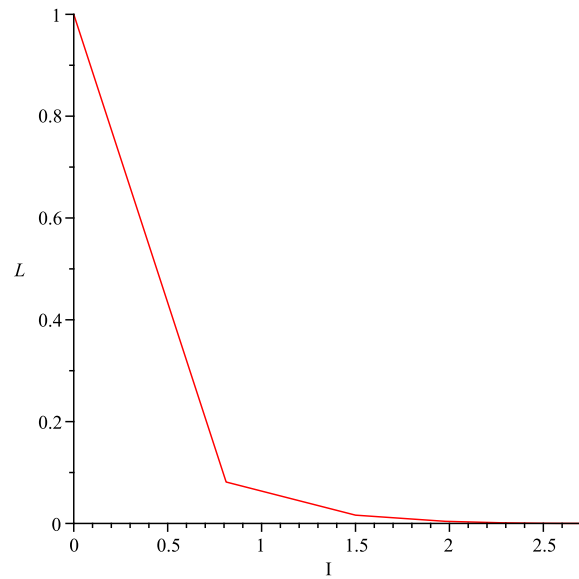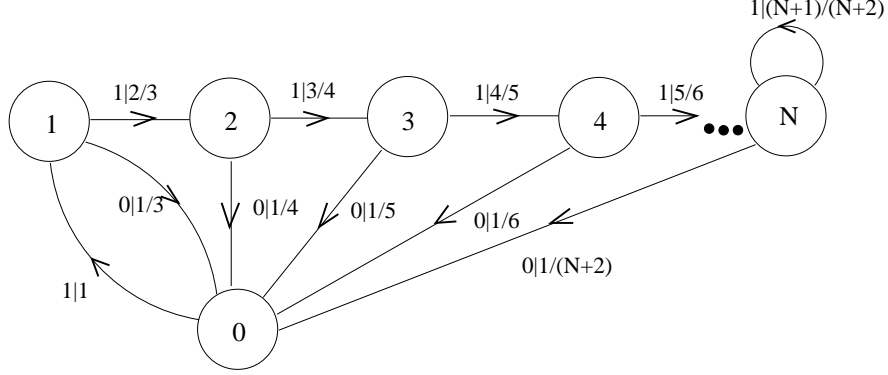FIG. 7. Sequential approximations to the exact ϵ-machine of Example A.

FIG. 8. Loss vs. information quasi-quantity for Example A.

TABLE IV. Quasi-information complexity C_γ for some values of γ for Example A.

γ	0.2	0.1	0.05	0.01	0
C_γ	0.8	1.2	1.4	1.6	2.7

FIG. 9. Exact ϵ-machine for Example B.

let the transition probabilities be $\Pr(i \to i+1) = \frac{i+1}{i+2}$, $\Pr(i \to 0) = \frac{1}{i+2}$ for $i = 1, \ldots N-1$, $\Pr(N \to N) = \frac{N+1}{N+2}$ and $\Pr(0 \to 1) = 1$. Just like in Example A, a transition from a state to another state with a numerically higher label or to itself is accompanied by an output of 1, and a transition from any state to state 0 outputs 0. Note that the number of states for this ϵ-machine has to be finite (but can be made arbitrarily large) in order for steady state probabilities to be positive. One can easily see that the steady state probabilities for this ϵ-machine are $P(0) = \frac{1}{2(\Psi(N+2)+\gamma_E-1/2)}$ and $P(i) = \frac{2}{i+1}P(0)$, where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function and $\gamma_E \approx 0.577$ is the Euler constant. Let the "partial" ϵ-machines representing incomplete information for this example have the same form as in the original example (see Fig. 10).

The probabilities shown in Fig. 10 have the following values:

$$q_0 = \frac{1}{2} \cdot \frac{1 + 2(\Psi(N+3) - \Psi(3))}{1 + \Psi(N+2) - \Psi(3)}$$

and

$$q_k = \frac{\Psi(N+3) - \Psi(k+2)}{\Psi(N+2) - \Psi(k+1)}$$

for $k = 1, 2, \ldots, N$. The steady state probabilities for the "partial" ϵ-machines can be computed in the same way as those for the exact one, and thus the information quasi-quantities equal to the corresponding Shannon entropies. On the other hand, the value of loss is found in a straightforward fashion using the general expression (13) and taking into account that the map $g \colon \mathcal{M}^* \to \mathcal{M}_k$ in this case has the form $g(j) = j$ if $j < k$ and $g(j) = k$ if $j \geq k$, for the "partial" ϵ-machine $\mathcal{M}_k$
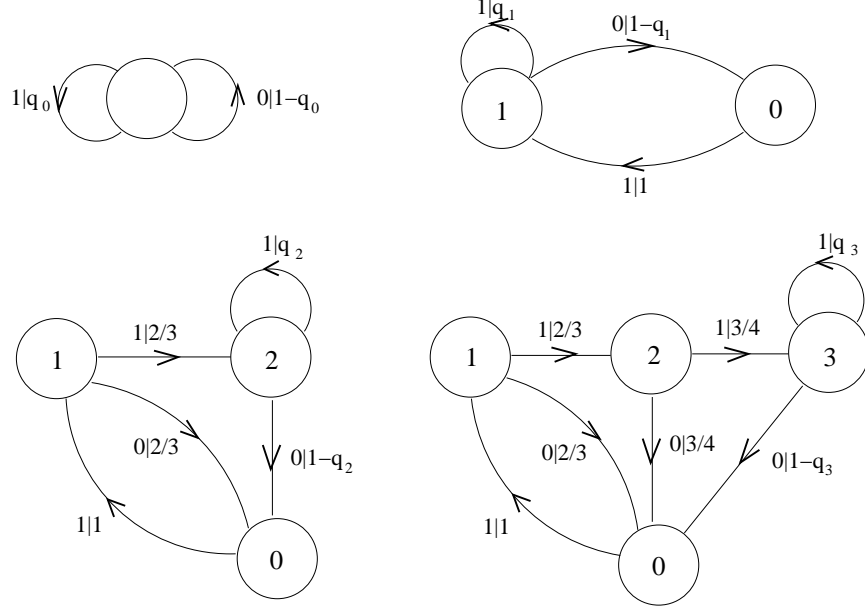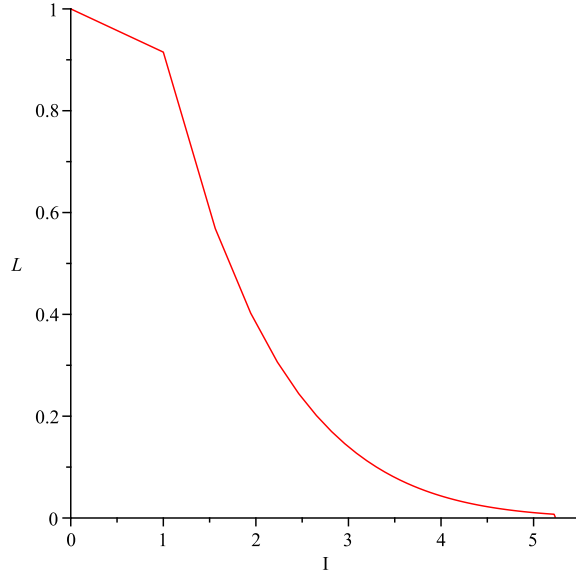
FIG. 10. Sequential approximations to the exact ϵ-machine of Example B.

FIG. 11. Loss vs. information quasi-quantity for Example B.

with $k + 1$ nodes. The resulting dependence of loss (in the convention $L_0 = L(\mathcal{M}_0) = 1$ and setting $N = 1000$) on the information quasi-quantity is shown in Fig. 11. The values of the quasi-information complexity C_γ for different values of γ can be found directly for this dependence. Some of these values are shown in Table V.

Now, consider an ϵ-machine shown in Fig. 12. We will refer to this example as Example C. As can be seen from the figure, this ϵ-machine has N states where N is a (large) even number, and the

TABLE V. Quasi-information complexity C_γ for some values of γ for Example B.

γ	0.2	0.1	0.05	0.01	0
C_γ	2.7	3.3	3.9	5.0	7.8

transition probabilities have the form $\Pr(i \to i+1) = \epsilon_i$, $\Pr(i \to i) = 1 - \epsilon_i$ for $i = 1, \ldots, N-1$, $\Pr(N \to 1) = \epsilon_N$, $\Pr(N \to N) = 1 - \epsilon_N$. Any transition to an even numbered state (either from an odd numbered state or from itself) outputs 1, and any transition to an odd numbered state outputs 0. It is easy to see that the steady state probabilities for this system have the form $P(i) = \frac{1-\epsilon_1}{1-\epsilon_i}P(1)$ for $i = 2, \ldots, N$. If one sets $\epsilon_i = \frac{1}{bi}$ where $b > 1$ is a real number, then a simple computation shows that $P(i) = B\frac{bi}{bi-1}$, $i = 1, \ldots, N$, where $B = \left(N + \frac{1}{b}(\Psi(N+1-\frac{1}{b}) - \Psi(1-\frac{1}{b}))\right)^{-1}$.

Let the incomplete information about the system be represented by the ϵ-machines shown in Fig. 13. The first of these is obtained, just like in the original example, by simply estimating the overall probabilities of 1 and 0 in the output assuming independence of output symbols (thus an absence of any "information processing" by the system). A simple calculation gives the following values for the probability p_1:

$$p_1 = \frac{1}{2} \cdot \frac{N + \frac{1}{b}\left(\Psi(\frac{1}{2}N + 1 - \frac{1}{2b}) - \Psi(1 - \frac{1}{2b})\right)}{N + \frac{1}{b}\left(\Psi(N + 1 - \frac{1}{b}) - \Psi(1 - \frac{1}{b})\right)},$$

and $p_0 = 1 - p_1$. The second ϵ-machine in Fig. 13 is obtained by mapping all odd numbered states of the exact ϵ-machine into the state O and all even numbered states – into the state E. The transition probabilities between states O and E are as shown in Fig. 13 where

$$p_0^O = \frac{\Psi(\frac{1}{2}N + 1 - \frac{b+1}{2b}) - \Psi(1 - \frac{b+1}{2b})}{Nb + \Psi(\frac{1}{2}N + 1 - \frac{b+1}{2b}) - \Psi(1 - \frac{b+1}{2b})},$$

$$p_1^E = \frac{\Psi(\frac{1}{2}N + 1 - \frac{1}{2b}) - \Psi(1 - \frac{1}{2b})}{Nb + \Psi(\frac{1}{2}N + 1 - \frac{1}{2b}) - \Psi(1 - \frac{1}{2b})},$$

and $p_1^O = 1 - p_0^O$, $p_0^E = 1 - p_1^E$, respectively. Any transition into state O is accompanied by an output of 0, and any transition into state E – by an output of 1.

If one sets the parameter b equal to 3 and $N = 1000$, the statistical complexity of the system calculates to be equal to 9.97. On the other hand, for the same values of parameters b and N, the loss corresponding to the two "partial" ϵ-machines shown in Fig. 13 is $L(\mathcal{M}_0) = 0.98$ and $L(\mathcal{M}_1) = 0.008$, with their information quasi-quantities (Shannon entropies) being equal to 0 and 1.0, respectively. It follows that, $C_{0.008} = 1.0$ for this system (or, rather, that 1.0 is a upper bound on $C_{0.008}$ as explained earlier).

We can now make some interesting observations.
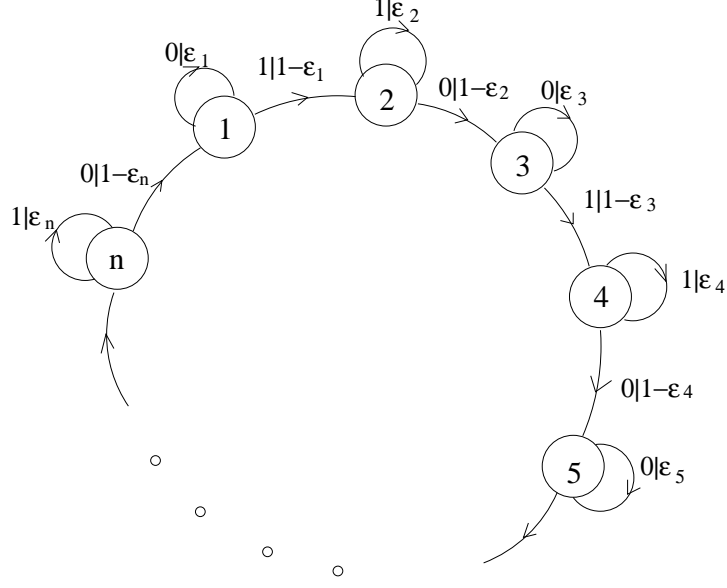
FIG. 12. Exact "circular" ϵ-machine for Example C.
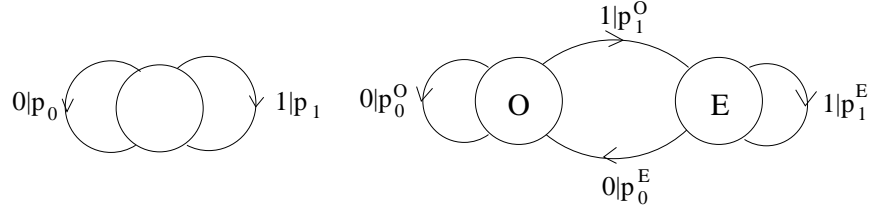
FIG. 13. First two approximations, $\mathcal{M}_0$ and $\mathcal{M}_1$, to the "circular" ϵ-machine of Example C.

- Comparing Examples A and B, we notice that the "full" quasi-information complexity C_0 (which is equal to statistical complexity) in Example B is significantly higher (7.8 vs 2.7) reflecting higher steady state probabilities of higher numbered states for Example B.

- The quasi-information complexity values for $\gamma > 0$ are also higher in the absolute sense but are about the same in the relative (measured as fractions of C_0) sense for Example B, compared to Example A. The latter observation can likely be explained by the two ϵ-machines sharing the same topology.

- If we now compare Examples A and C, we can see that the statistical complexity of the system in Example C is significantly higher than that in Example A (10.0 vs 2.7), again reflecting higher steady state probabilities of larger number of states in Example C.

- On the other hand, the quantity $C_{0.01}$ is actually *lower* for the "circular" system in Example C (1.0 vs. 1.6). Informally speaking, this means that if a relative 1% difference in prediction

quality is immaterial to the agent, the system in Example C would appear less complex than that in Example A.

- In this particular case, the latter phenomenon has a simple explanation. Indeed, if the agent predicts the next bit based on the hypothesis of i.i.d. distribution (which corresponds to the agent using the simplest ϵ-machine $\mathcal{M}_0$ possessing no "structure") the resulting error is a lot larger than that afforded by the knowledge of the 'full' ϵ-machine (complete information). In fact, the bits appear to be almost "unpredictable" (as the estimated probability of 1 is close to 0.5). The situation changes drastically if the agent is able to get access to $\mathcal{M}_1$ with two causal states, i.e. able to realize that in this system 0 is followed by 1 and vice versa with high probability. The knowledge of $\mathcal{M}_1$ allows the agent to predict nearly as well as if he had access to the full ϵ-machine $\mathcal{M}^*$ with 1000 causal states. Put slightly differently, once the realization is made that 0 and 1 symbols "mostly" alternate in this system, the rest of information just introduces small corrections to this dominant behavior.

- Summarizing, one can say that to an agent not concerned (for whatever reason) with small (1% or less) "residual" losses, the system from Example C would appear as genuinely less complex than that in Example A, in spite of a much larger statistical complexity in Example C.

Informally speaking, an agent observing the system output in Example C would quickly observe that 0's and 1's mostly alternate but sometimes shorts strings of the same symbol are observed. This observation would let the agent conclude that the system can be reasonably accurately described by an ϵ-machine of the form $\mathcal{M}_1$ presented earlier. It would take the agent, generally speaking, a lot more effort to discern finer details: namely that length of these short string of the same symbol tend to change just a little as time passes. The discovery of these fine details though would improve the agent's predictive ability only very slightly. One could say that, for this particular system, it is relatively easy to understand its behavior "almost fully" but a lot more difficult to fill in the remaining small details. The initially very steep C_γ curve that flattens out drastically reflects that.

V. CONCLUSION

The main contributions of this article can be briefly summarized as follows.

- The notion of information-related complexity was extended from systems described by

stochastic processes to more general tasks (performed by a rational agent), including optimization and decision making problems with uncertainty about input data.

- A general problem oriented definition of complexity was proposed and it was shown that it reduces to a generalization of statistical complexity in the case the system output is described by a discrete-time stochastic process and the agent's task is that of prediction .

- The degree of task completion was explicitly considered and it was argued that the information-related complexity should be properly understood relative to the extent of task completion, or, equivalently, properly defined loss. In other words, the information-related complexity for the given task is inherently a real-valued function, and not just a real number.

- It was argued, based on results of [10, 11], that the minimum information quantity necessary to complete the task to the given degree should be really interpreted as appropriate quasi-quantity that takes into account the degree of difficulty of obtaining the corresponding information in the given environment. The traditional information quantity measures (such as Shannon entropy) obtain from the appropriate quasi-quantity in the particular case of maximum symmetry, when either all different "pieces" of information are identical (with respect to the degree of difficulty of their acquisition by the agent) or their differences are unknown.

Let's discuss some of these points in an informal way. A more complex system is the one that "takes a lot of work" to fully understand for a rational agent, and, moreover, a complex system still takes a lot of work to understand "almost fully". On the contrary, a system that is "easy" to understand to a large degree but very hard to "get the rest" (a system with for which the C_γ curve is initially very steep but flattens out at close to zero loss) would appear to be of low complexity to all but "very discerning" agents. One might ask: what about the true "objective" complexity of a system that's independent (being objective) on all and any agents. We believe that this question might be more difficult and deeper than it appears. One possible answer is that it depends on the point of view and the level of inquiry. The same system can be complex from one point of view and much less so from another. Additionally, it can be complex at high level of detail but much less complex at "lower resolution". For instance, in natural sciences, a system's ability to "innovate", or spontaneously create new quality could be due to various levels of its structure detail. Hence, its appropriate complexity would also depend on that. Generally speaking, it is our opinion that

the science of complexity still has more questions than answers, and it would be of use, for the sake of clarity, to make some implicitly used or assumed aspects of the overall problem – explicit. This is one of the main goals of the present article.

---

[1] A. N. Kolmogorov, Problems of Information Transmission **1**, 4 (1958).

[2] A. N. Kolmogorov, IEEE Trans. Inf. Theory **14**, 662 (1968).

[3] T. Hogg and B. A. Huberman, "Order, complexity, and disorder," (1985), palo Alto preprint.

[4] P. Grassberger, International Journal of Theoretical Physics **25**, 907 (1986).

[5] C. H. Bennett, in *Emerging Syntheses in Science*, edited by D. Pines (Santa Fe Institute, Santa Fe, NM, 1985) pp. 297-313.

[6] C. H. Bennett, Foundations of Physics **16**, 585 (1986).

[7] J. P. Crutchfiled and K. Young, Physical Review Letters **63**, 105 (1989).

[8] M. Koppel, Complex Systems **1**, 1087 (1987).

[9] C. R. Shalizi and J. P. Crutchfiled, Journal of Statistical Physics **104**, 817 (2001).

[10] E. Perevalov and D. Grace, "Towards the full information chain theory: question difficulty," (2012), submitted to Physical Review E, arXiv:1212.2696[physics.data-an].

[11] E. Perevalov and D. Grace, "Towards the full information chain theory: answer depth and source models," (2012), submitted to Physical Review E, arXiv:1212.2693[physics.data-an].

[12] R. T. Cox, Am. J. Phys. **14**, 1 (1946).

[13] R. T. Cox, *The Algebra of Probable Inference* (Johns Hopkins Press, Baltimore, 1961).

[14] R. T. Cox, in *The Maximum Entropy Formalism*, edited by R. Levine and M. Tribus (MIT Press, Cambridge, MA, 1979) pp. 119-167.

[15] K. H. Knuth, Neurocomputing **67**, 245274 (2005).

[16] K. H. Knuth, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Sao Paolo, Brazil* (2008) pp. 24–31.

[17] J. P. Crutchfiled, Physica D **75**, 11 (1994).

[18] Here an ϵ-machine is understood as simply a function on a Markov chain, without a direct connection to any dynamic system or stochastic process.

[19] Strictly speaking, according to definitions of [9], the "partial ϵ-machines" used here to describe incomplete information are not ϵ-machines in the proper sense since their states are not *causal states* (i.e. equivalence classes of past histories characterized by the same distributions of futures, or *morphs*) , but rather simply *effective states* (i.e. equivalence classes of past histories that in general do not possess that property). Nevertheless, we refer to them as ("partial") ϵ-machines just to avoid introducing additional notions.

[20] It is easy to see that such state S'' is unique, due to deterministic property of ϵ-machine $\mathcal{M}$.